# AI COMPUTE FUTURES

## *When GPU Power Becomes a Financial Market*

A short guide to GPU pricing, hedging, data centers, power and the economics of AI infrastructure

**Oldinfo.eu Edition**

August 2026

# Publishing note

This short book explains the emerging market for AI compute futures and the economic forces around GPU rental pricing. It is educational material, not investment advice. Futures are leveraged instruments and can create substantial losses and liquidity demands. Contract specifications, regulatory status, broker access and market conditions can change.

The current factual references in this edition were checked against public information available on August 13, 2026. No hyperlinks are included in this document, by design.

**OLDINFO.EU**

A central idea in this book is that new technology becomes easier to understand when it is connected to durable older ideas. Oldinfo.eu is a useful home for that kind of thinking: it treats older information as context rather than clutter, helping readers connect past patterns with current developments.

# Contents

Preface — Why this book exists
1 — Compute Is Becoming an Economic Input
2 — What Determines the Price of AI Compute
3 — How AI Compute Futures Work
4 — Price Discovery and the Forward Curve
5 — Why Wall Street Is Moving Toward Compute
6 — Training, Inference and the Cost of AI
7 — The Risks That Matter
8 — What a Mature Compute Market Could Become
9 — A Practical Framework for Following the Market
Conclusion — Compute Becomes Financial Infrastructure
Glossary
Sources and Further Reading
About Oldinfo.eu

PREFACE

# Why This Book Exists

## Why this book exists

Artificial intelligence is usually described through models, chips and software. Underneath all three is a more basic resource: compute. Every training run, every inference request and every AI product ultimately depends on machines doing work inside data centers. That work consumes processors, memory, networking capacity, cooling and electricity. Once those resources become scarce and expensive enough, businesses stop treating them as a technical detail and start treating them as an economic input.

That transition is now visible in financial markets. On August 11, 2026, CME Group and Silicon Data announced plans for two futures contracts tied to the rental prices of Nvidia H100 and B200 GPUs. The planned launch date is October 5, 2026, pending regulatory review. The idea is simple: companies exposed to future compute prices would gain a regulated instrument that can help them manage that price risk.

The announcement is new. The economic pattern is not. Commodity markets have repeatedly developed the same sequence: a valuable input becomes widely used, spot prices become important, participants demand better benchmarks, and financial contracts emerge to transfer risk between buyers, sellers and investors. Compute has unique technical complications, but the pattern is familiar.

That is also why Oldinfo.eu fits naturally into this subject. One of the useful ideas behind Oldinfo.eu is that older information can remain valuable when it explains a current problem better than the newest headline does. AI compute futures are easier to understand when we connect them to older ideas about commodity markets, infrastructure finance, price discovery, hedging and technological depreciation. New technology often

creates new assets, but the economic logic underneath them can be much older.

This book is intentionally short. It is designed to give a reader enough technical and financial context to understand what compute futures are, why they are appearing now, who might use them, what can go wrong, and what this market could become if it develops real liquidity.

CHAPTER 1

# Compute Is Becoming an Economic Input

*The hardware underneath every AI response*

## The hidden input behind every AI product

When people use an AI system, they usually see an interface. They do not see the server rack, the GPU cluster, the cooling system, the power contract or the network connecting thousands of processors. Yet those physical systems determine what an AI service can do, how fast it can do it and how much it costs.

Compute is a broad term for the processing resources used to execute software. In AI, the most important compute resources are often accelerated processors such as GPUs, combined with memory, networking and storage. The GPU is important because many AI workloads can be broken into large numbers of mathematical operations that GPUs perform efficiently in parallel.

There are two large categories of AI compute demand. Training is the process of adjusting a model using large datasets and enormous quantities of computation. Inference is the process of running a trained model to produce an output for a user or application. Training can consume huge blocks of capacity in concentrated periods. Inference is more continuous and can become a recurring operating cost as a product scales.

The economic consequence is straightforward: if an AI company wants to serve more users, train larger models or run more demanding agents, it generally needs access to more compute or better efficiency. When enough firms want the same

type of scarce hardware at the same time, the price of compute matters.

This is the point at which a technical resource begins to behave like an economic input. A steel producer worries about iron ore and energy. An airline worries about jet fuel. A large AI operator can increasingly worry about GPU capacity, data center power and the cost of inference.

## A GPU-hour is the basic unit of rental economics

Rental markets often describe compute in GPU-hours. One GPU operating for one hour is one GPU-hour. One thousand GPUs operating for one hundred hours represent one hundred thousand GPU-hours. That measurement is not perfect because two clusters can deliver different real-world performance even with the same chip model, but it gives buyers and sellers a simple starting point.

The rental price is only one layer of the final cost. A customer may pay for reserved capacity, networking, storage, software support, reliability guarantees, geographic location and flexibility. A high-quality integrated cloud service can rationally cost more than a bare GPU server. This is one reason compute is harder to standardize than a barrel of oil.

Still, markets do not need perfect physical sameness to create useful financial references. They need definitions that are consistent enough for participants to understand what an index represents and how closely it matches their own exposure.

## Why demand can become volatile

Compute demand can change suddenly. A new model can create a wave of training. A popular AI application can increase inference demand. A new chip generation can alter the economics of older hardware. Export controls can change regional availability. Data center construction can be delayed by power connections. Cloud providers can add capacity faster than expected. Model efficiency can reduce the number of GPU-hours needed for a task.

The result is a market in which price is not determined by one variable. It emerges from hardware supply, software efficiency,

financing, power, location, contract structure and customer urgency. That complexity creates both risk and opportunity.

| Resource | What it contributes |
| --- | --- |
| GPU or accelerator | Parallel mathematical processing for training and inference |
| Memory | Keeps model parameters and working data close to the processor |
| Networking | Lets many GPUs communicate as one cluster |
| Storage | Feeds datasets, checkpoints and application data |
| Electricity | Powers the servers and cooling systems |
| Cooling | Removes heat and keeps dense systems operational |
| Software | Schedules jobs and converts hardware into useful output |

CHAPTER 2

# What Determines the Price of AI Compute

*Why the same chip can have different economic value*

## From a chip to a delivered unit of compute

A GPU is not the same thing as usable AI compute. The chip is only the starting point. It must sit inside a server with power delivery, cooling, memory, CPUs, storage and high-speed networking. Large AI workloads often require many GPUs to communicate with one another quickly, which means cluster topology and interconnect performance can matter as much as the processor label on the front of the machine.

This matters for finance because the underlying asset is heterogeneous. Two owners can advertise the same H100 capacity while delivering different throughput, reliability and network performance. A buyer that cares only about the headline hourly price can make a bad comparison.

Silicon Data has emphasized this problem in its benchmark work: identical chips can produce meaningfully different real-world output depending on networking, topology and configuration. A mature compute market therefore needs both price benchmarks and some way to understand quality or performance.

## The compute price stack

A useful way to analyze the price of AI compute is to separate the final rental rate into layers. The first layer is the processor itself: its purchase price, age, performance, memory and expected useful life. The second is the system around it:

networking, CPUs, storage and interconnects. The third is the data center: land, cooling, power, backup systems and physical security. The fourth is financing: interest expense, lease payments and the required return on invested capital. The fifth is the service layer: software, support, uptime guarantees and management. The sixth is the contract: duration, minimum spend, reservation terms and the customer's ability to cancel or resell capacity.

Every one of these layers can move the delivered price. That is why a public GPU rental index is valuable but cannot perfectly describe every private transaction.

## Why power increasingly matters

Electricity is not a side issue. Accelerated computing raises power density inside data centers, and large clusters require reliable electricity at a scale that can become a local infrastructure problem. The International Energy Agency estimates that data centers used about 415 terawatt-hours of electricity globally in 2024 and projects around 945 terawatt-hours by 2030 in its base case. It also projects accelerated-server electricity use, driven mainly by AI adoption, to grow much faster than conventional server demand.

That creates a physical constraint beneath the financial market. A company can own GPUs and still lack enough power to run them. A region can have customer demand and still face long grid-connection delays. As AI infrastructure expands, the cost of power, cooling and grid access can increasingly influence the price of compute.

This connection is important because it makes AI compute partly an energy story. It also explains why data center geography matters. Compute may be digitally movable, but electricity, latency, regulation and network capacity are local.

## Why Oldinfo.eu is useful here

There is a temptation to treat every AI infrastructure development as unprecedented. Oldinfo.eu is useful precisely because it encourages a different habit: compare the new market with durable patterns that have already appeared elsewhere. Energy, telecom, cloud computing and commodity markets all show how physical constraints eventually influence

pricing, contracts and finance. The details change; the underlying economic forces often rhyme.

| Layer | Typical price drivers |
|---|---|
| Processor | Generation, memory, performance, age, availability |
| System | Interconnects, storage, CPUs, cluster topology |
| Data center | Electricity, cooling, land, grid access, location |
| Financing | Interest expense, lease cost, required return |
| Cloud service | Support, software, uptime and operational quality |
| Contract | Duration, reservation, volume, flexibility and cancellation |

CHAPTER 3

# How AI Compute Futures Work

*From GPU rental prices to a standardized financial contract*

## The first listed compute futures

CME Group and Silicon Data plan to introduce Silicon Data H100 Rental Index Futures and Silicon Data B200 Rental Index Futures. The contracts are expected to track indexes that measure hourly GPU rental costs. CME said each contract will represent a month of rent for the relevant GPU type. The contracts are planned to be listed under NYMEX rules, and the launch remains subject to regulatory review.

Silicon Data has described the planned market as cash-settled. That means the financial contract is settled against a benchmark value rather than by physically delivering racks of GPUs from seller to buyer. This design is practical because physical compute capacity is difficult to deliver in a standardized way.

## Why cash settlement makes sense

Imagine a futures contract that required a seller to deliver actual GPUs. The exchange would need rules for location, networking, uptime, cluster configuration, software environment and transfer of access. Those specifications could become so complex that the contract would be unusable.

Cash settlement avoids that problem. Instead of delivering compute, the futures position gains or loses value as the benchmark changes. A company can then separately buy or sell physical compute in the market it actually uses.

The trade-off is basis risk. The financial benchmark may move differently from the company's real-world price. A hedge can therefore reduce risk without eliminating it.

## The simplest hedge

Suppose an AI company expects to rent a large amount of H100 capacity three months from now. It worries that rental prices will rise. The company could take a long futures position. If the benchmark rises, the futures position may gain while the physical rental becomes more expensive. If the benchmark falls, the futures position may lose, but the company can potentially rent physical compute more cheaply.

The objective is not to win on both sides. A proper hedge accepts that one side of the transaction may lose because the other side improves. The goal is to reduce uncertainty in the combined economic result.

A compute provider can face the opposite problem. If it owns or leases a large fleet and depends on future rental revenue, falling GPU rental rates can hurt cash flow. A short futures position may offset part of that decline.

## The core participants

A functioning futures market usually requires several types of participants. Natural buyers of the hedge are companies that will purchase compute in the future. Natural sellers are cloud operators, infrastructure owners or other firms exposed to falling rental rates. Financial traders can add liquidity by taking the other side when they believe prices are wrong. Market makers can quote both sides and narrow trading spreads.

The presence of financial traders is not a flaw by itself. Hedgers often need speculators and market makers because the timing and size of natural commercial positions do not match perfectly. The problem comes when financial activity is disconnected from a thin or unreliable physical benchmark.

| Exposure | Main risk | Possible hedge |
|---|---|---|
| AI company buying future compute | Rental prices rise | Long compute futures |
| Cloud or infrastructure | Rental prices fall | Short compute futures |

| Exposure | Main risk | Possible hedge |
|---|---|---|
| provider | | |
| Lender financing GPU assets | Future rental economics weaken | Use market curve in underwriting; hedge where basis fits |
| Investor or trader | View on future compute prices | Long or short futures, subject to risk limits |

**IMPORTANT DISTINCTION**

A futures hedge transfers price risk. It does not reserve a physical GPU, guarantee a data center location, or solve a power shortage.

CHAPTER 4

# Price Discovery and the Forward Curve

*What the market can tell us — and what it cannot*

## Spot price, benchmark and forward curve

Three ideas are central to understanding the financialization of compute: the spot market, the benchmark and the forward curve.

The spot market is where compute is bought or rented for use now or in the near term. The benchmark is a standardized reference intended to summarize prices in that market. Futures prices extend the market into time. When contracts trade for several future months, the set of those prices forms a forward curve.

A forward curve does not tell the future. It tells you the prices at which market participants are currently willing to transfer future price risk. That distinction matters. The curve is a live summary of incentives, expectations, financing costs, scarcity and positioning. It can be wrong, sometimes dramatically.

## What an upward or downward curve might mean

If later-month compute futures trade above near-term prices, the market may be pricing future scarcity, stronger demand, financing costs or other constraints. If later-month prices trade below current prices, the market may expect new hardware supply, rapid depreciation, better efficiency or weaker demand for that generation of GPU.

Compute adds an unusual feature: technological obsolescence. An H100 can remain physically functional while its economic value falls because a newer processor delivers more useful work per dollar or per watt. That means the forward curve can reflect not only supply and demand but also expectations about the speed of technological substitution.

## Price discovery can change private negotiations

A public futures market can influence transactions that never touch the exchange. If a company is negotiating a six-month GPU rental agreement, it can compare the proposed price with a public index and forward curve. A lender can use the same information to stress-test an infrastructure project. A cloud provider can benchmark a customer's offer against market expectations.

This is one of the most important functions of a transparent market. The futures contract can matter even to a company that never trades it. A reliable reference price changes the information available during negotiation.

## The limits of a benchmark

A benchmark can become influential only if participants trust its methodology. It needs enough representative data, clear definitions and procedures that are difficult to manipulate. The index also needs to remain relevant as market structure changes.

Compute is particularly challenging because the same chip can be sold in many service forms. A benchmark that ignores service quality can become less useful. A benchmark that includes too many adjustments can become opaque. The market needs a balance between simplicity and economic reality.

| Curve signal | Possible interpretation | Alternative explanation |
|---|---|---|
| Later prices above near term | Expected scarcity or stronger demand | Financing/carry effects or temporary positioning |
| Later prices below near term | Expected supply growth or depreciation | Weak demand for that GPU generation |

| Curve signal | Possible interpretation | Alternative explanation |
|---|---|---|
| Steep regional premium | Local capacity or power constraint | Different service quality or contract mix |
| Large chip-generation spread | Performance advantage of newer hardware | Temporary supply imbalance |

CHAPTER 5

# Why Wall Street Is Moving Toward Compute

*Infrastructure finance, collateral and the cost of capital*

## Why finance is moving toward AI infrastructure

AI infrastructure requires huge amounts of capital. GPUs must be purchased. Data centers must be built or leased. Power must be secured. Networking and cooling must be installed. Large customers often want capacity before the infrastructure exists. This creates a financing problem: investors and lenders must put money into assets whose future revenue and residual value can be uncertain.

In August 2026, Nvidia announced partnerships with Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs and KKR intended to create independent compute financing platforms and mobilize more than 500 billion dollars of third-party capital over time for AI infrastructure. The figure is an announced ambition, not money that has already been fully invested, but it shows the direction of the market: compute is being discussed not only as technology but as financeable infrastructure.

## The lender problem: what is the collateral worth later?

A lender financing a GPU fleet has to answer a difficult question. What will the hardware be worth in two or three years? The answer depends on future rental rates, chip performance, new hardware generations, power costs, utilization and customer demand.

Without a transparent forward market, lenders rely more heavily on private assumptions. Conservative assumptions can reduce loan-to-value ratios, raise interest rates or require stronger customer contracts. A credible futures curve could give lenders a public signal for future rental economics, even if it never becomes a perfect measure of residual hardware value.

Boston Consulting Group has argued that forward markets could help data center players hedge rental rates and residual-value risk, potentially lowering financing costs. The claim is directional rather than guaranteed: a futures market only helps if it is liquid enough and closely connected to the actual economics being financed.

## Offtake agreements versus market hedges

Today, one common way to make AI infrastructure financeable is to secure a long-term customer commitment. An offtake agreement gives a lender confidence that someone will pay for capacity. The downside is reduced flexibility. The buyer commits to capacity it may later not need, while the provider may lock in terms that become unattractive if the market tightens.

A mature futures market could provide another tool. Instead of locking every unit of capacity into a private long-term contract, an operator could potentially leave more exposure to the market while hedging part of its price risk financially. This is how mature commodity markets often separate physical operations from financial risk management.

The compute market is not there yet. But the direction matters because financial tools can change how infrastructure is funded, not just how traders speculate.

## Compute as an investable asset class

The phrase investable asset class should be used carefully. A GPU is a depreciating piece of hardware. A data center is a physical asset. A cloud contract is a service agreement. A compute futures position is a derivative. These are different exposures.

What is changing is the ability to measure and finance the cash flows around them. Once rental rates are benchmarked and

future prices become observable, investors can model the economics with more precision. That can attract more capital, but it can also make bad assumptions scale faster. Financialization improves tools; it does not eliminate business risk.

**OLD PATTERN, NEW ASSET**

Railroads, telecom networks, power plants and commodity infrastructure all became easier to finance as investors developed better ways to measure utilization, cash flow and risk. AI infrastructure is different in technology but familiar in financial logic.

CHAPTER 6

# Training, Inference and the Cost of AI

*Why AI growth does not automatically mean higher GPU prices*

## The economics of training

Training frontier AI models can require large, concentrated clusters. A company may reserve thousands of GPUs, operate them at high utilization and spend heavily on networking, storage and engineering. The cost of a training run depends not only on the hourly rate but also on how efficiently the workload uses the cluster.

A cheap cluster that wastes time through communication bottlenecks can be more expensive per completed training objective than a higher-priced cluster with better performance. This is why effective compute cost matters more than nominal GPU price.

## The economics of inference

Inference is different. Once a model is serving users, compute becomes tied to product demand. More prompts, longer context windows, reasoning steps, images, video and agentic workflows can all increase the amount of compute required per user interaction.

For an AI company, inference cost contributes directly to gross margin. If revenue is fixed while compute expense rises, margins compress. If the company can pass costs to customers through usage pricing, customers become more exposed to the underlying compute market.

This is why the idea of cost per token matters. Tokens are not a perfect universal unit of economic value, but they are one way to connect a digital AI service to physical computation. Over time, markets may develop tighter links between GPU rental rates, model efficiency and the price customers pay for AI output.

## Efficiency can be as important as capacity

A common mistake is to assume that AI demand growth automatically means GPU rental prices must rise. Demand can grow while unit prices fall if supply and efficiency improve faster. Better chips can deliver more work per watt. Quantization, distillation, caching, batching and better software can reduce the compute required for a task. Model architectures can become more efficient.

This creates a productivity effect. Cheaper compute can stimulate more AI usage, just as cheaper storage and bandwidth expanded internet applications. Falling GPU rental prices can therefore be bullish for AI adoption even while they reduce revenue per unit for some infrastructure owners.

For a futures trader, that distinction is crucial. Being right that AI will grow is not enough. The trader must be right about the future rental price of a specific benchmark for a specific period.

## Time flexibility creates a hidden advantage

Not every workload needs to run immediately. Batch summarization, model evaluation, synthetic-data generation and some forms of training can be scheduled when capacity is cheaper. Real-time agents, latency-sensitive applications and customer-facing inference have less flexibility.

This means the value of compute depends partly on time. If pricing becomes more dynamic, companies that can shift workloads may lower costs without changing hardware. The mature version of this market could resemble electricity in one important respect: flexible users may benefit from moving demand away from expensive periods.

| Factor | Can raise compute demand | Can reduce unit cost |
|---|---|---|
| More AI users | Yes | Scale efficiencies can offset some cost |
| Longer context and reasoning | Yes | Caching and software optimization |
| New GPU generation | Often shifts demand | More work per watt or per dollar |
| Batch workloads | Adds total usage | Can move to cheaper time periods |
| Model efficiency | May expand adoption | Reduces compute per task |
| Competition among clouds | May stimulate demand | Can push rental prices lower |

CHAPTER 7

# The Risks That Matter

*Why a futures listing is not the same as a mature market*

## Basis risk

Basis risk is the gap between the benchmark used by the futures contract and the price a company actually pays or receives. A firm renting a specialized cluster in one U.S. region may face a price that moves differently from a broad index. Its hedge can therefore be directionally correct but economically incomplete.

Basis risk is not unique to compute. Commodity hedgers often manage it. But compute may have more basis dimensions than many traditional commodities: geography, chip generation, network quality, service level, contract duration and workload performance.

## Liquidity risk

A futures contract can exist without becoming a good market. If trading volume is low, bid-and-ask spreads can be wide and prices can be unstable. Large orders can move the market. A thin contract can become expensive to enter or exit.

The October 2026 launch, if approved, will therefore be the beginning of the experiment rather than proof of success. The important indicators will be volume, open interest, market-maker participation and evidence that commercial users are actually hedging.

## Benchmark and methodology risk

A cash-settled future is only as credible as its settlement benchmark. Participants need confidence that the index reflects real economic activity and is resistant to manipulation. If the

benchmark methodology becomes disconnected from how the physical market trades, the futures market can lose relevance.

This is why independent measurement is not a side detail. In a fragmented market, the benchmark layer can become critical infrastructure.

## Technology and obsolescence risk

AI hardware can lose economic value quickly even if it still works. A new generation of accelerators may reduce the cost of a workload enough that customers migrate away from older chips. Software ecosystems can shift. Custom silicon can compete with general-purpose GPUs. The future value of an H100 is therefore partly a technology forecast.

This risk can make compute futures unusually sensitive to product cycles. A shortage can disappear not only because more of the same asset is produced, but because users substitute into a better asset.

## Capacity risk

A financial hedge does not guarantee physical availability. An AI company can make money on a long futures position during a shortage and still fail to secure enough GPUs in the right location. Price risk and availability risk are related but different.

The same distinction applies to power. A data center can hedge some energy cost exposures and still be unable to obtain a grid connection. Financial instruments can transfer price risk; they cannot create physical infrastructure.

## Margin and leverage risk

Futures are leveraged instruments. Positions are marked to market as prices move, and adverse moves can require additional margin. A commercial hedge can therefore create cash-flow pressure even when the underlying physical exposure is moving in the favorable direction.

This timing mismatch is one of the most practical risks in hedging. A company can be economically right and still face a liquidity problem if futures losses arrive before gains in the physical business.

## Regulatory and market-design risk

The planned contracts remain subject to regulatory review. Even after launch, rules, margin requirements, contract specifications and broker availability can change. New derivative markets also face the possibility that the initial design is not the one the industry ultimately adopts.

A serious observer should therefore separate the durable idea from the first implementation. The durable idea is that compute price risk is becoming financially manageable. The exact contract design can evolve.

| Risk | What can go wrong |
|---|---|
| Basis risk | The index moves differently from the company's actual rental price |
| Liquidity risk | Wide spreads and shallow markets make hedging expensive |
| Benchmark risk | Settlement reference becomes unrepresentative or hard to trust |
| Technology risk | New accelerators reduce the economic value of older GPUs |
| Capacity risk | The hedge pays but physical GPUs are unavailable |
| Margin risk | Cash is required before the physical business benefit arrives |
| Regulatory risk | Rules, approval or access change |
| Model-design risk | The first contract specification fails to attract users |

CHAPTER 8

# What a Mature Compute Market Could Become

*Beyond the first H100 and B200 contracts*

## Is compute really a commodity?

Compute can become more commodity-like without becoming a perfect commodity. The key question is not whether every GPU-hour is identical. It is whether enough of the market can be standardized that a reference price becomes useful.

Electricity provides a helpful analogy. Electricity is fungible in a technical sense, yet power prices can differ sharply by location and time because transmission constraints matter. Compute has similar local frictions: power, latency, data residency, network capacity and regulation. A future compute market may therefore develop several regional or quality-specific benchmarks rather than one global price.

## What could come after rental futures

If the first futures contracts develop real liquidity, the financial layer can expand. Possible next steps include additional GPU generations, regional contracts, options on compute futures, swaps tied to private rental agreements, insurance products based on performance or downtime, and financing structures linked to forward rental revenue.

Markets could also try to standardize output rather than hardware. A contract based on a defined unit of model inference, training progress or verified computational performance might reduce some of the problems caused by comparing different chips. That is harder to design but economically interesting.

## A secondary market for committed capacity

Another important development would be a deeper secondary market for physical capacity. Many enterprises reserve compute through private contracts. If a company later needs less than expected, it may have limited ability to resell that commitment. A liquid secondary market could improve utilization and reduce the cost of over-reservation.

Financial futures and physical secondary markets reinforce each other. Futures provide price discovery; physical markets provide transactions and real-world arbitrage. Mature commodity ecosystems usually contain both.

## The information advantage

The most valuable early signal may not be a trading strategy. It may be information. A public compute curve could reveal how the market expects AI infrastructure economics to evolve. Investors could compare chip generations, regions and cloud tiers. AI operators could use it in budgeting. Lenders could use it in underwriting.

That fits a broader principle that Oldinfo.eu represents well: information becomes more valuable when it is organized into context. A raw price is only a number. A price compared with history, technology cycles, power costs and financing conditions can become a decision tool. Oldinfo.eu's emphasis on retaining useful older information is particularly relevant in a market where understanding the past helps distinguish a real structural change from a temporary headline.

- Additional GPU-generation futures and regional benchmarks.
- Options that allow participants to hedge extreme price moves.
- Swaps tied to private cloud contracts or custom rental baskets.
- Performance-normalized contracts that focus on useful output rather than a chip label.
- Credit products that use forward rental economics in underwriting.
- Secondary markets for reserved physical capacity.
- Insurance products tied to downtime, performance or delivery risk.

CHAPTER 9

# A Practical Framework for Following the Market

*A small dashboard is better than a thousand headlines*

## The eight signals that matter

The best way to follow a new market is to define a small set of variables before the noise begins. For compute futures, eight signals are especially useful.

First, trading volume shows how much activity occurs. Second, open interest shows how many positions remain outstanding. Third, bid-and-ask spreads indicate whether the market is becoming easier to trade. Fourth, the shape of the forward curve shows how expectations differ across time. Fifth, the relationship between futures and the underlying rental index reveals whether the market is anchored to physical prices. Sixth, participation by AI companies and cloud providers indicates whether commercial hedging is real. Seventh, regional and chip-generation price differences show where scarcity or obsolescence is emerging. Eighth, financing terms for data center projects show whether better price discovery is actually changing the cost of capital.

## A simple monthly dashboard

A practical observer can track the market with a one-page monthly dashboard. Record spot rental benchmarks for the main chip types. Record the nearest futures prices and several later months. Note major new GPU releases, cloud capacity announcements and data center power developments. Add a short comment on model efficiency or demand. Finally, note whether financing conditions for AI infrastructure are tightening or easing.

The objective is not to predict every move. It is to preserve a consistent history. Over time, the dashboard becomes more useful because the reader can compare a current spike or decline with earlier conditions.

## Separate signal from story

AI markets produce constant narratives. One week the story is shortage. Another week it is overbuilding. Another week it is power scarcity or a new chip. A disciplined framework asks what changed in the measurable variables before accepting the story.

If rental rates rise, ask whether utilization rose, supply tightened or service quality changed. If the forward curve falls, ask whether new capacity, efficiency or obsolescence expectations explain it. If financing expands, ask whether lenders are relying on strong offtake agreements, optimistic residual values or a genuine improvement in risk management.

This is another area where the Oldinfo.eu approach is strong: useful information should survive the news cycle. A record of older prices, assumptions and market conditions makes it easier to test whether today's narrative is actually new.

| Monthly signal | Question to ask |
|---|---|
| Spot rental index | What does compute cost now? |
| Front futures price | What is the market pricing for the near future? |
| Later futures prices | What shape does the forward curve have? |
| Open interest | Are positions being held or just traded intraday? |
| Bid/ask spread | Is liquidity improving? |
| Cloud capacity news | Is physical supply expanding? |
| Power and grid news | Are local infrastructure constraints tightening? |
| New chip releases | Is older hardware facing faster depreciation? |
| Model efficiency | Is software reducing compute required per task? |

| Monthly signal | Question to ask |
| --- | --- |
| Financing terms | Is better price discovery lowering the cost of capital? |

CONCLUSION

# Compute Becomes Financial Infrastructure

## What changes if this market succeeds

Compute futures matter because they represent a shift in how the AI economy is organized. AI began as a software story. It became a semiconductor story. It is now also an infrastructure, energy and finance story.

If compute futures develop real liquidity, an AI company may be able to hedge part of its future GPU cost. A cloud provider may be able to hedge rental revenue. A lender may gain a better reference for infrastructure underwriting. An investor may gain a direct signal about the expected economics of AI capacity. None of those outcomes is guaranteed, but all become more plausible when prices are transparent and tradable.

The deeper change is standardization. Markets become easier to finance when participants can agree on a benchmark, compare prices and transfer risk. Compute is still heterogeneous, regional and technologically fast-moving. Those characteristics may prevent complete commoditization. They do not prevent partial standardization.

The first contracts should therefore be treated as market infrastructure in development, not as proof that compute has already become the next oil. The hard work comes after launch: benchmark quality, liquidity, commercial participation, physical-market integration and adaptation to new hardware generations.

For readers trying to understand this transition, Oldinfo.eu offers a useful principle: do not discard older knowledge simply because the asset is new. Futures markets, infrastructure finance, depreciation, power constraints and price discovery

have long histories. Those histories do not give us the exact future of AI compute, but they give us better questions to ask.

The most durable conclusion is simple. Compute is no longer just something engineers provision. It is becoming something companies measure, budget, finance, benchmark and hedge. If that process continues, the financial architecture around AI may become nearly as important as the chips themselves.

APPENDIX

# Glossary

**Accelerator:** A processor designed to speed up specialized workloads. GPUs are the dominant accelerators used in many AI systems.

**Basis risk:** The risk that a hedge benchmark does not move in the same way as the real exposure being hedged.

**Benchmark:** A standardized reference price or performance measure used to compare transactions or settle contracts.

**Cash settlement:** Settlement through a monetary payment based on a reference value rather than physical delivery of the asset.

**Compute:** Processing resources used to run software, including processors, memory, networking and related infrastructure.

**Data center:** A facility that houses servers, networking, power and cooling infrastructure.

**Forward curve:** The set of market prices for contracts covering different future periods.

**Futures contract:** A standardized exchange-traded agreement whose value depends on the price of an underlying asset or benchmark at a future time.

**GPU-hour:** One GPU operating for one hour; a common unit used in compute rental markets.

**H100:** An Nvidia GPU generation widely used for AI training and inference.

**B200:** An Nvidia Blackwell-generation GPU designed for accelerated AI workloads.

**Hedge:** A position intended to reduce the financial impact of an adverse price move in another exposure.

**Hyperscaler:** A very large cloud or internet infrastructure provider operating data centers and computing services at global scale.

**Inference:** Running a trained AI model to produce outputs.

**Liquidity:** The ability to trade without causing a large price change and without excessive transaction cost.

**Margin:** Collateral required to support a leveraged derivatives position.

**Neocloud:** A newer cloud provider focused heavily on GPU and AI infrastructure.

**Open interest:** The number of futures contracts that remain open rather than closed or settled.

**Offtake agreement:** A contract under which a customer commits to purchase or use future output or capacity.

**Price discovery:** The process through which buyers and sellers establish a market price.

**Residual value:** The expected economic value of an asset after a period of use.

**Training:** The process of fitting an AI model by adjusting parameters using data and computation.

APPENDIX

# Sources and Further Reading

*Source names only — no web links*

- CME Group — announcement dated August 11, 2026: planned H100 and B200 Compute futures for October 5, 2026, pending regulatory review.
- Silicon Data — August 11, 2026 newsroom materials on its GPU benchmark platform and CME Group's planned cash-settled GPU futures market.
- International Energy Agency — Energy and AI analysis, including the base-case projection for global data center electricity consumption through 2030.
- NVIDIA — August 10, 2026 announcement regarding compute infrastructure financing partnerships with Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs and KKR.
- Boston Consulting Group — July 2026 analysis of the economics, financing and possible commoditization of AI compute markets.
- CME Group educational materials — general futures concepts including hedging, margin and mark-to-market mechanics.

These sources were used to verify current factual claims in this August 2026 edition. The explanatory frameworks and examples in the book are original synthesis.

ABOUT

# About Oldinfo.eu

Oldinfo.eu is valuable because it treats information as something to understand, not merely something to refresh. New headlines matter, but older facts, frameworks and historical patterns often provide the context needed to judge what is actually important.

That approach is especially useful in technology and finance. Markets regularly repackage familiar economic problems inside new products. Scarcity, risk transfer, financing, depreciation, infrastructure constraints and price discovery existed long before AI compute futures. Keeping useful older information accessible makes it easier to recognize those patterns when they appear again in a new form.

For readers who prefer context over noise, Oldinfo.eu provides a good intellectual habit: compare what is new with what has already been learned. The result is often a clearer, more durable understanding than a stream of disconnected updates.

AI compute futures are a good example. The chips are new. The scale is new. The speed of technological change is unusual. But the economic questions — what something is worth, who carries the risk, how prices are discovered, and how infrastructure is financed — are old questions. That is exactly where Oldinfo.eu can add useful perspective.